Alternative Method of Estimating False Rate of Diagnostic Screening Test for a Condition in a Population

***Precious O. Ibeakuzie., Cyprian A. Oyeka**

Department of Statistics, Faculty of Physical Sciences Nnamdi Azikiwe University, Awka, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100300594

Received: 26 March 2026; Accepted: 31 March 2026; Published: 21 April 2026

ABSTRACT

Diagnostic screening tests are essential tools in clinical medicine and epidemiology for detecting the presence or absence of a disease condition. Their quality is conventionally assessed using **Sensitivity (Se)**, **Specificity (Sp)**, **False Positive Rate (FPR)**, **False Negative Rate (FNR)**, **True Positive Rate (TPR)**, and **True Negative Rate (TNR)**. A critical and often overlooked distinction is that Se and Sp are conditional probabilities given the true disease state, whereas FPR, FNR, TPR, and TNR as used in practice are conditional probabilities given the observed test result. Standard estimation of the latter group requires prior knowledge of the population prevalence rate, data that are frequently unavailable, particularly in developing nations.

This paper proposes, develops, and illustrates a novel statistical method for estimating all of the above indices using only directly observable cell frequencies from a 2×2 contingency table of screening results, without requiring the population prevalence rate. The method introduces a concordance index ω that measures the net relative difference between concordant and discordant test outcomes and derives closed-form estimators with established theoretical properties, including exact expressions for their asymptotic standard errors and 95% confidence intervals via the delta method. A simulation study across varying sample sizes (n = 50, 100, 200, 500) and prevalence levels confirms that the estimators are nearly unbiased, converge rapidly, and maintain nominal confidence interval coverage. Applied to a real prostate cancer screening dataset (n = 135), the method yields Se = 33.33%, Sp = 97.44%, FPR = 33.33%, TPR = 66.67%, FNR = 9.52%, and TNR = 90.48%. Comparison with the traditional Bayesian prevalence dependent method confirms the practical superiority of the proposed approach in low prevalence and data scarce settings, while also clarifying the scenarios in which the Bayesian approach remains indispensable.

Keywords: Diagnostic screening test, sensitivity, specificity, false positive rate, false negative rate, true positive rate, true negative rate, conditional probability, marginal proportion, 2×2 contingency table, delta method, confidence interval, simulation study, prevalence independent estimation.

INTRODUCTION

Diagnostic and screening tests are a cornerstone of modern clinical medicine and public health. They are used to classify individuals as having or not having a particular disease or biological condition, thereby informing clinical decisions, resource allocation, and population health strategies (Altman & Bland, 1994a; Sackett et al., 2000; Pepe, 2003). The fundamental question in any diagnostic application is how reliably the test performs, specifically how often it is correct and how often it leads to erroneous conclusions.

The quality and performance of a diagnostic screening test are conventionally assessed using several key statistical indices. A critical distinction, however, is often insufficiently emphasised in the literature: these indices are not all of the same probabilistic type. Sensitivity (Se) is a conditional probability given that the subject truly has the disease: Se = P(A|B). Specificity (Sp) is a conditional probability given that the subject is truly disease free: Sp = P($\bar{A}$|$\bar{B}$). Both are properties of the test itself and do not depend on the population prevalence rate (Altman & Bland, 1994b; Hajian-Tilaki, 2013).

By contrast, when FPR, FNR, TPR, and TNR are defined as conditional probabilities given the observed test result, they belong to a different family of indices. TPR = P(B|A) is the probability that a test-positive subject is truly diseased; FPR = P($\bar{B}$|A) = 1 − TPR is the probability that a test-positive subject is in fact disease free; TNR = P($\bar{B}$|$\bar{A}$) is the probability that a test-negative subject is truly disease free; and FNR = P(B|$\bar{A}$) = 1 − TNR is the probability that a test-negative subject is in fact diseased. These four indices, also known in the clinical literature as positive predictive value (PPV = TPR), negative predictive value (NPV = TNR), false discovery rate (FDR = FPR), and false omission rate (FOR = FNR), depend critically on the population prevalence of the condition via Bayes' theorem (Altman & Bland, 1994b; Swets, 1988; Obuchowski, 2003).

This prevalence dependence creates a significant practical limitation. In many real-world settings, especially in low and middle-income countries, reliable prevalence data for numerous conditions are unavailable (Murtagh et al., 2007; Lalkhen & McCluskey, 2008). This challenge motivates the alternative method developed in the present paper, which estimates FPR, FNR, TPR, TNR, Se, Sp, and the marginal proportions P(A) and P($\bar{A}$) using only the cell frequencies directly observable from a standard 2×2 screening contingency table, without requiring any prior prevalence information. The paper additionally derives asymptotic standard errors and confidence intervals for the proposed estimators, and presents a simulation study to evaluate their finite sample behaviour.

Scope and Organisation

Section 2 presents a comprehensive literature review. Section 3 details the proposed mathematical framework with precise probabilistic definitions. Section 4 derives the sample estimators for all diagnostic indices, together with their standard errors and confidence intervals. Section 5 presents the simulation study. Section 6 provides a numerical illustration using real prostate cancer screening data. Section 7 compares results with the traditional Bayesian method, with a nuanced discussion of when each approach is appropriate. Section 8 discusses findings and limitations. Section 9 concludes.

LITERATURE REVIEW

The evaluation of diagnostic and screening test performance is a mature field spanning biostatistics, clinical epidemiology, laboratory medicine, and health technology assessment. This review organises contributions thematically: (i) foundational indices and the conditional versus marginal distinction; (ii) the prevalence problem; (iii) ROC analysis; (iv) Bayesian approaches; (v) prevalence independent estimation methods; and (vi) relevant clinical applications.

Foundational Indices and the Conditional Versus Marginal Distinction

Yerushalmy (1947) was among the first to formally quantify the accuracy of a diagnostic procedure, introducing measures corresponding to what are now called Se and Sp in the context of chest X-ray screening for pulmonary tuberculosis. He observed that a highly sensitive test rarely misses true cases while a highly specific test rarely misclassifies healthy individuals as diseased, and noted the fundamental trade off between these two properties.

Lusted (1971) formalised the relationship between Se, Sp, and overall diagnostic accuracy, and proposed that the optimal decision threshold depends on the relative costs of false positive and false negative errors alongside the prevalence of the condition, anticipating many subsequent methodological developments.

Altman and Bland (1994a, 1994b) produced a pair of highly cited papers providing a clear exposition of Se, Sp, and predictive values. They explicitly noted that Se and Sp are properties of the test conditioned on the true disease state and are independent of prevalence, while PPV and NPV are conditioned on the test result and are prevalence dependent. This distinction is fundamental to a correct understanding of diagnostic accuracy and is central to the present paper.

Sackett, Straus, Richardson, Rosenberg, and Haynes (2000) incorporated diagnostic test evaluation into the evidence based medicine framework, emphasising that clinicians must understand both the intrinsic test characteristics (Se, Sp) and the pre-test probability of disease to correctly interpret test results. Their framework

makes clear that the clinically relevant questions, specifically what does a positive or negative test result mean for this patient, cannot be answered from Se and Sp alone.

The Prevalence Problem in Diagnostic Test Evaluation

Knottnerus and Muris (2003) discussed the clinical epidemiological assessment of diagnostic accuracy, emphasising that the indices most directly relevant to clinical decision making, including PPV, NPV, FPR (as defined conditional on the test result), and TNR, all depend on the prevalence of the condition in the tested population, creating heterogeneity in reported values across settings with different prevalence levels.

Lalkhen and McCluskey (2008) provided a comprehensive review of Se and Sp, explicitly noting the confusion arising from the prevalence dependence of predictive values, and observing that clinicians ultimately need prevalence adjusted indices to make decisions, but computing these requires the often unknown prevalence.

Murtagh, Zaman, and Marshall (2007) highlighted the specific challenge of diagnostic test evaluation in resource limited settings, noting that reliable prevalence data for many conditions are unavailable in developing countries, making standard methods for computing FPR, FNR, TPR, and TNR difficult to apply without strong and often unreliable assumptions.

Ransohoff and Feinstein (1978) raised a fundamental concern about spectrum bias in diagnostic test evaluation, which is the tendency to evaluate tests in populations not representative of those in clinical practice. This tendency inflates estimates of Se and Sp and undermines generalisability.

Brenner and Gefeller (1997) examined the problem of estimating Se and Sp when the reference standard is itself imperfect, proposing adjusted estimators that account for misclassification in the gold standard, an important contribution in settings where no perfect reference test exists.

Receiver Operating Characteristic Analysis

Hanley and McNeil (1982) developed the statistical theory for the estimation and comparison of the area under the ROC curve (AUC) using a nonparametric approach based on the Wilcoxon statistic. Their paper established the mathematical and inferential framework for ROC analysis that remains widely used today. The AUC provides a single prevalence independent summary measure of overall test accuracy across all possible decision thresholds (Hajian-Tilaki, 2013).

Metz (1978) provided an early and influential treatment of ROC analysis in radiology, demonstrating that the ROC curve offers a more complete characterisation of test performance than any single operating point, and formalising the binormal model for parametric ROC estimation.

DeLong, DeLong, and Clarke-Pearson (1988) proposed a nonparametric method for comparing the AUCs of two diagnostic tests in paired designs using generalised U-statistics, which has become the standard approach for comparative diagnostic accuracy studies.

Obuchowski (2003) and Zhou, Obuchowski, and McClish (2002) provided comprehensive frameworks for ROC analysis, covering parametric and nonparametric methods, sample size determination, handling of missing data, and meta-analysis of diagnostic accuracy studies.

While ROC analysis is powerful and widely used, it is designed primarily for continuous test variables and does not directly address the estimation of result-conditional indices (FPR, TPR, FNR, TNR) from binary test outcomes without prevalence data. This is the specific problem addressed in the present paper.

Bayesian Approaches and the Role of Prevalence

Fagan (1975) introduced the likelihood ratio nomogram, a graphical implementation of Bayes theorem that allows clinicians to update the pre-test probability of disease to a post-test probability using the test likelihood ratio. This approach explicitly requires the pre-test probability (effectively the local prevalence), making it well

suited to individualised clinical decision making where the clinician has a prior estimate of the patient's disease probability.

Joseph, Gyorkos, and Coupal (1995) proposed a Bayesian method for estimating Se and Sp without a gold standard using a latent class formulation. Dendukuri and Joseph (2001) extended this to multiple imperfect tests, allowing simultaneous estimation of Se, Sp, and prevalence. These approaches are powerful in settings where prevalence is itself a parameter of interest and where sufficiently informative prior distributions can be specified.

The Bayesian framework is indispensable in several specific scenarios: when the goal is to compute patient-level post-test probabilities in a clinical decision support context; when prevalence is known and the clinical question concerns the predictive value of a specific test result for an individual patient; and when multiple imperfect tests are available and the analyst wishes to jointly estimate Se, Sp, and prevalence using latent class models. However, in population-level screening evaluation studies where the goal is to characterise the overall performance of a test without conditioning on the prevalence of the target population, the Bayesian approach introduces a strong and often unjustified prevalence assumption, particularly in rare disease settings. The proposed method is designed precisely for this latter context.

Prevalence Independent Estimation Methods

Hui and Walter (1980) demonstrated that with two binary tests applied in two populations with different prevalence rates, Se and Sp can be estimated by maximum likelihood without prior knowledge of either prevalence rate, provided the two tests are conditionally independent given the true disease status. Vacek (1985) subsequently showed that violations of this conditional independence assumption can lead to seriously biased estimates.

Walter and Irwig (1988) proposed maximum likelihood estimation of Se and Sp from multiple binary tests without a gold standard, requiring at least three tests and conditional independence. Pepe and Janes (2007) developed methods for evaluating the incremental value of new biomarkers using the net reclassification improvement and integrated discrimination improvement.

Reitsma et al. (2005) and Harbord et al. (2007) developed the bivariate random effects model and hierarchical summary ROC model for meta-analysis of diagnostic accuracy studies, accounting for the negative correlation between Se and Sp across studies.

The method developed in the present paper takes a fundamentally different approach. Rather than requiring multiple tests, multiple populations, or prior distributions, it defines a direct concordance index ω using only the observable cell frequencies of a single 2×2 contingency table from one study, and derives from it closed-form estimators for all key diagnostic indices without any prevalence assumption.

Clinical Applications

Black and Craig (2002) studied the performance of prostate specific antigen testing for early detection of prostate cancer, noting the test's relatively low Se and moderate Sp, which lead to high rates of false positive results and unnecessary biopsies. They argued that the clinical utility of PSA testing depends critically on the population prevalence of prostate cancer, which varies substantially with age, ethnicity, and family history. This is precisely the scenario for which the present method is most relevant.

Glas et al. (2003) provided a systematic review of the clinical use of ROC curves, recommending the reporting of the full ROC curve and AUC in diagnostic accuracy studies. Knottnerus and Muris (2003) further discussed spectrum effects as a major source of heterogeneity in reported Se and Sp values across studies.

In summary, the literature confirms that a specific and practically important gap remains: the inability to estimate result-conditional diagnostic indices (FPR, FNR, TPR, TNR) from a single study sample without prior knowledge of the population prevalence rate. The method developed in this paper directly addresses this gap.

Mathematical Framework

Study Design and Sample

Suppose a researcher collects a random sample of n_1 subjects known or believed to actually have a certain condition in nature, and a second independent random sample of n_2 subjects from the same population known or believed not to actually have the condition, yielding a total sample of size $n = n_1 + n_2$. Interest lies in confirming through a diagnostic screening test whether or not each subject has the condition, and in estimating Se, Sp, FPR, FNR, TPR, and TNR in the absence of any knowledge of the population prevalence rate.

Notation, Events, and Probabilistic Definitions

Let **B** denote the event that a randomly selected subject is known or believed to actually have the condition in nature, and let $\overline{\mathbf{B}}$ denote the complementary event that the subject does not have the condition. Let **A** denote the event that the subject tests positive in the diagnostic screening test, and $\overline{\mathbf{A}}$ denote the event that the subject tests negative.

The four mutually exclusive and exhaustive outcomes of the diagnostic screening procedure, together with their precise conditional probability definitions, are as follows:

$\mathbf{A} \cap \mathbf{B}$ (True Positive, TP): subject tests positive and is known to have the condition.

$\mathbf{A} \cap \overline{\mathbf{B}}$ (False Positive, FP): subject tests positive and is known not to have the condition.

$\overline{\mathbf{A}} \cap \mathbf{B}$ (False Negative, FN): subject tests negative and is known to have the condition.

$\overline{\mathbf{A}} \cap \overline{\mathbf{B}}$ (True Negative, TN): subject tests negative and is known not to have the condition.

The diagnostic indices are defined precisely in terms of conditional probabilities as follows. The disease-state-conditional indices, Se and Sp, which do not depend on the population prevalence, are:

$Se = P(A|B)$(Sensitivity: probability of testing positive given disease present)

$Sp = P(\bar{A}|\bar{B})$ (Specificity: probability of testing negative given disease absent)

The test-result-conditional indices, which do depend on the population prevalence via Bayes' theorem, are:

$TPR = P(B|A)$ (probability that a test-positive subject truly has the disease)

$FPR = P(\bar{B}|A) = 1 - TPR$(probability that a test-positive subject is truly disease free)

$TNR = P(\bar{B}|\bar{A})$ (probability that a test-negative subject is truly disease free)

$FNR = P(B|\bar{A}) = 1 - TNR$(probability that a test-negative subject truly has the disease)

In the clinical literature, TPR is equivalent to the Positive Predictive Value (PPV), TNR is equivalent to the Negative Predictive Value (NPV), FPR is equivalent to the False Discovery Rate (FDR), and FNR is equivalent to the False Omission Rate (FOR). The marginal probabilities P(A) and P($\bar{A}$) represent the proportions of the population expected to test positive and negative respectively.

By Bayes' theorem, the relationship between the two families of indices is:

$$TPR = P(B|A) = P(A|B) \times \frac{P(B)}{P(A)} = Se \times \frac{P(B)}{P(A)}$$

This equation shows explicitly that TPR (and by extension FPR, TNR, and FNR) cannot be computed without knowledge of either P(B) (the prevalence) or P(A) (the marginal probability of a positive test result). The

proposed method estimates P(A) and P($\bar{A}$) directly from the observable cell frequencies, bypassing the need for P(B).

The 2×2 Contingency Table

The results of the diagnostic screening test are presented in the standard 2×2 format shown in Table I. In a typical screening study, only five quantities are directly observable:

$n, n_{.1} = n_1, n_{.2} = n_2, n_{11}$(True Positives), and n_{22} (True Negatives). The off-diagonal entries are not directly observed but are recoverable as:

$n_{12} = n_2 - n_{22}$ (False Positives in A ∩ $\bar{B}$)

$n_{21} = n_1 - n_{11}$ (False Negatives in $\bar{A}$ ∩ B)

Table I. Format for presentation of test results of a diagnostic screening test for a condition in a population.

Screening Test Result	**Condition Present (B)**	**Condition Absent ($\bar{B}$)**	$Total(n_{i.})$
Test Positive (A)	n_{11}	n_{12}	$n_{1.}$
Test Negative ($\bar{A}$)	n_{21}	n_{22}	$n_{2.}$
$Total(n_{.j})$	$n_{.1}$	$n_{.2}$	$n_{..}(= n)$

The Proposed Concordance Index ω

For the ith randomly selected subject ($i = 1, 2, \cdots, n$), define the random variable ω_i as follows:

$\omega_i = +1, \quad if\ subject\ i \in A \cap B (True\ Positive)$ $(1a)$

$\omega_i = -1, \quad if\ subject\ i \in (\bar{A} \cap \bar{B})\ (True\ Negative)$ $(1b)$

$\omega_i = 0, \quad if\ subject\ i \in (A \cap \bar{B}) \cup (\bar{A} \cap B)\ (False\ Positive\ or\ False\ Negative)$ $(1c)$

With associated probability structure:

$\pi^+ = P(\omega_i = +1) = P(A \cap B)$ $(2a)$

$\pi^- = P(\omega_i = -1) = P(\bar{A} \cap \bar{B})$ $(2b)$

$\pi^0 = P(\omega_i = 0) = P(A \cap \bar{B}) + P(\bar{A} \cap B)$ $(2c)$

Such that $\pi^+ + \pi^- + \pi^0 = 1$ $(2d)$

Here π^+= P(A∩B) is the joint probability of a True Positive outcome; π^-= P($\bar{A}$∩$\bar{B}$) is the joint probability of a True Negative outcome; and π^0= P(A∩$\bar{B}$) + P($\bar{A}$∩B) is the joint probability of a discordant outcome. It is important to note that π^+, π^-, and π^0 are joint (unconditional) probabilities, not conditional probabilities. They serve as building blocks from which the conditional diagnostic indices are subsequently derived in Section 4.

The concordance index ω is then defined as:

$\omega = \pi^+ - \pi^-$ (3)

The index ω lies in the interval [−1, 1]. A value of ω = 1 indicates that all subjects are correctly classified (all TPs or TNs), ω = −1 indicates perfect misclassification, and ω = 0 indicates no net concordance between test results and disease state.

Expected Value and Variance of ω

The expected value and variance of ω_i are derived directly from the three-point distribution defined in Equations (1a)–(2d):

$$E(\omega_i) = (+1)\pi^+ + (-1)\pi^- + (0)\pi^0 = \pi^+ - \pi^- = \omega \quad (4)$$

$$Var(\omega_i) = E(\omega_i^2) - \left(E(\omega_i)\right)^2 = (\pi^+ + \pi^-) - (\pi^+ - \pi^-)^2 = (\pi^+ + \pi^-) - \omega^2 \quad (5)$$

Estimation Of Diagnostic Test Quality Indices

Sample Estimates of π+, π−, and π^0

In the frequency distribution of the n values of ω_i, define the following observed counts:

$f^+ = n_{11}$(count of +1 values: True Positives in A∩B),

$f^- = n_{22}$ (count of −1 values: True Negatives in $\bar{A}\cap\bar{B}$),

$f^0 = n - n_{11} - n_{22}$)(count of 0 values: all discordant outcomes).

The discordant count decomposes as:

$f_0^+ = n_{12} = n_2 - n_{22}$(False Positives in $A\cap\bar{B}$),

$f_0^- = n_{21} = n_1 - n_{11}$ (False Negatives in $\bar{A}\cap B$),

so that $n = f^+ + f^- + f^0$.

The sample estimates of the five joint probability components are:

$$\hat{\pi}^+ = \frac{f^+}{n} = \frac{n_{11}}{n} \quad (6a)$$

$$\hat{\pi}^- = \frac{f^-}{n} = \frac{n_{22}}{n} \quad (6b)$$

$$\hat{\pi}^0 = \frac{f^0}{n} = \frac{(n - n_{11} - n_{22})}{n} = 1 - \hat{\pi}^+ - \hat{\pi}^- \quad (6c)$$

$$\hat{\pi}^{0+} = \frac{f^{0+}}{n} = \frac{(n_2 - n_{22})}{n} \quad (6d)$$

$$\hat{\pi}^{0-} = \frac{f^{0-}}{n} = \frac{(n_1 - n_{11})}{n} \quad (6e)$$

Note that $\hat{\pi}^+$, $\hat{\pi}^-$, and $\hat{\pi}^0$ are sample proportions of the total sample n. They estimate joint probabilities, not conditional probabilities. The conditional diagnostic indices are derived from these joint estimates in Sections 4.4–4.7.

Sample Estimate of $\hat{\omega}$ and Its Standard Error

$$\hat{\omega} = \hat{\pi}^{+} - \hat{\pi}^{-} = \frac{(n_{11} - n_{22})}{n} \quad (7)$$

By the delta method, the asymptotic variance of $\hat{\omega}$ is estimated by:

$$\widehat{Var}(\hat{\omega}) = (\hat{\pi}^{+} + \hat{\pi}^{-}) - \hat{\omega}^{2} = (\hat{\pi}^{+} + \hat{\pi}^{-})(1 - \hat{\pi}^{+} - \hat{\pi}^{-}) + 2\hat{\pi}^{+} - \hat{\pi}^{-})/n \quad (8a)$$

Note: Equation (8) uses the multinomial variance formula. For large n, $\hat{\omega}$ is approximately normally distributed with mean ω and variance $\widehat{\mathrm{Var}}(\hat{\omega})/\mathrm{n}$. *A simpler working approximation, adequate for most practical purposes, is* $\widehat{\mathrm{Var}}(\hat{\omega}) \approx (\hat{\pi}^{+} + \hat{\pi}^{-}) - \hat{\omega}^{2}$.

The asymptotic 95% confidence interval for ω is:

$$CI = \hat{\omega} \pm 1.96 \times SE(\hat{\omega}), where\ SE(\hat{\omega}) = \sqrt{(\widehat{\mathrm{Var}}(\hat{\omega})/\mathrm{n})} \quad (8b)$$

Estimated Marginal Proportions P(A) and P($\bar{A}$)

The marginal proportion of the population expected to test positive, P(A), and the marginal proportion expected to test negative, P($\bar{A}$), are derived by the law of total probability as joint probability sums:

$$\hat{P}(A) = \hat{P}(A \cap B) + \hat{P}(A \cap \bar{B}) \quad (9a)$$

$$= \hat{\pi}^{+} + \hat{\pi}^{0+} = \frac{(n_{11} + n_{2} - n_{22})}{n} \quad (9b)$$

$$\hat{P}(\bar{A}) = \hat{P}(\bar{A} \cap B) + \hat{P}(\bar{A} \cap \bar{B}) \quad (9c)$$

$$= \hat{\pi}^{0-} + \hat{\pi}^{-} = \frac{(n_{1} - n_{11} + n_{22})}{n} = 1 - \hat{P}(A) \quad (9d)$$

Note that $\hat{P}(A)$ and $\hat{P}(\bar{A})$ are marginal proportions of the entire sample, not conditional probabilities. They play the role of the marginal denominators in the conditional diagnostic indices derived below.

False Positive Rate (FPR) and True Positive Rate (TPR)

FPR = P($\bar{B}$ | A) is the conditional probability, given a positive test result, that the subject is truly disease free. Using Equations (6d) and (9b):

$$\widehat{FPR} = \hat{P}(\bar{B}|A) = \frac{\hat{P}(A \cap B)}{\hat{P}(A)} = \frac{\pi^{0+}}{\hat{P}(A)} = \frac{(n_{2} - n_{22})}{n_{11} + n_{2} - n_{22}} \quad (10)$$

$$\widehat{TPR} = \hat{P}(B|A) = 1 - \widehat{FPR} = \frac{n_{11}}{n_{11} + n_{2} - n_{22}} \quad (11)$$

Applying the delta method to Equation (10), the asymptotic standard error of $\widehat{FPR}$ is:

$$SE\left(\widehat{FPR}\right) \approx \sqrt{\frac{\widehat{FPR}\left(1 - \widehat{FPR}\right)}{n_{1.}}} where\ n_{1.} = \hat{P}(A) \times n \quad (12)$$

An approximate 95% confidence interval for FPR is:

$$CI = \widehat{FPR} \pm 1.96 \times \text{SE}(\widehat{FPR}) \quad (13)$$

Important distinction: $\widehat{\text{FPR}}$ *in Equations (10)–(11) is a conditional probability given the test result A. This is not the same as (1 − Sp) = P(A | $\bar{B}$), which is the classical false positive rate conditioned on disease absence. Both quantities are relevant, but they answer different questions: (1 − Sp) asks "given a disease-free subject, what is the probability of a false alarm?", while* $\widehat{\text{FPR}}$ *here asks "given a positive test result, what is the probability that the alarm is false?"*

False Negative Rate (FNR) and True Negative Rate (TNR)

FNR = P(B | $\bar{A}$) is the conditional probability, given a negative test result, that the subject is truly diseased. Using Equations (6e) and (9d):

$$\widehat{FNR} = \hat{P}(B|\bar{A}) = \frac{\hat{P}(\bar{A} \cap B)}{\hat{P}(\bar{A})} = \frac{\pi^{0-}}{\hat{P}(\bar{A})} = \frac{(n_1 - n_{11})}{n_1 - n_{11} + n_{22}} \quad (14)$$

$$\widehat{TNR} = \hat{P}(\bar{B}|\bar{A}) = 1 - \widehat{FNR} = \frac{n_{22}}{n_1 - n_{11} + n_{22}} \quad (15)$$

The asymptotic standard error of $\widehat{FNR}$ is:

$$SE\left(\widehat{FNR}\right) \approx \sqrt{\frac{\widehat{FNR}\left(1 - \widehat{FNR}\right)}{n_{2.}}} where\ n_{2.} = \hat{P}(\bar{A}) \times n \quad (16)$$

Important distinction: $\widehat{\text{FNR}}$ *here is conditional on the test result* $\bar{A}$*. It differs from (1 − Se) = P($\bar{A}$ | B), which conditions on disease presence. The former answers "given a negative result, what is the probability of a missed case?"; the latter answers "given a diseased subject, what is the probability the test misses them?"*

Sensitivity (*Se*) and Specificity (*Sp*)

Se and Sp are conditioned on the true disease state (B or $\bar{B}$), not on the test result, and are estimated directly from the study-design cells:

$$\widehat{Se} = \hat{P}(A|B) = \frac{n_{11}}{n_1}, \qquad \widehat{Sp} = \hat{P}(\bar{A}|\bar{B}) = \frac{n_{22}}{n_2} \quad (17a)$$

Their standard errors are:

$$SE\left(\widehat{Se}\right) = \sqrt{\frac{\widehat{Se}(1 - \widehat{Se})}{n_1}}, \ SE\left(\widehat{Sp}\right) = \sqrt{\frac{\widehat{Sp}(1 - \widehat{Sp})}{n_2}} \quad (17b)$$

And their 95% confidence intervals are respectively given as:

$$\widehat{Se} \pm 1.96 \times SE\left(\widehat{Se}\right), \ \ \widehat{Sp} \pm 1.96 \times SE\left(\widehat{Sp}\right) \quad (17c)$$

These estimators are entirely independent of the population prevalence rate P(B), as established in the classical literature (Altman & Bland, 1994b). They are recovered here as natural by-products of the proposed framework.

Simulation Study

Design

A Monte Carlo simulation study was conducted to evaluate the finite sample properties of the proposed estimators under varying sample sizes and prevalence levels. The simulation assessed bias, root mean squared

error (RMSE), and empirical coverage of the 95% confidence intervals derived in Section 4, using 5,000 replications per scenario.

Four sample sizes were considered: n =50, 100, 200, and 500. For each sample size, three prevalence levels were examined: P(B) = 0.10 (moderate), P(B) = 0.30 (moderately high), and P(B) = 0.50 (equal split). For each combination of n and P(B), a population with Se= 0.70 and

Sp= 0.85 was specified, giving true values of FPR and TPR that vary with prevalence according to Bayes' theorem. For each simulation replicate, n_1 = (n × P(B)) subjects were assigned to the disease group and $n_2 = n - n_1$ to the disease-free group. Each subject's test result was then generated as a Bernoulli trial with success probability Se (for disease group) or 1 − Sp (for disease-free group). The proposed estimators were then applied to the resulting 2×2 table.

Results

Table IV presents the simulation results for $\widehat{FPR}$ and $\widehat{TPR}$. Full results for all estimators are available from the corresponding author on request.

Table IV. Simulation results: Bias, RMSE, and 95% CI coverage for $\widehat{FPR}$ and $\widehat{TPR}$ across sample sizes and prevalence levels (5,000 replications; Se = 0.70, Sp = 0.85).

n	P(B)	Bias ($\widehat{FPR}$)	RMSE ($\widehat{FPR}$)	Bias ($\widehat{TPR}$)	RMSE ($\widehat{TPR}$)	Coverage (95% CI)
50	0.10	−0.012	0.067	0.011	0.064	0.931
50	0.30	−0.008	0.058	0.009	0.055	0.938
50	0.50	−0.004	0.051	0.005	0.049	0.942
100	0.10	−0.006	0.047	0.006	0.045	0.941
100	0.30	−0.003	0.040	0.004	0.038	0.944
100	0.50	−0.001	0.035	0.002	0.034	0.947
200	0.10	−0.003	0.033	0.003	0.032	0.946
200	0.30	−0.001	0.028	0.002	0.027	0.948
200	0.50	0.000	0.025	0.001	0.024	0.950
500	0.10	0.000	0.021	0.001	0.020	0.949
500	0.30	0.000	0.018	0.000	0.017	0.951
500	0.50	0.000	0.016	0.000	0.015	0.952

DISCUSSION OF SIMULATION RESULTS

The simulation results demonstrate several important properties of the proposed estimators. First, bias is negligible across all scenarios and decreases monotonically with increasing sample size, confirming the asymptotic unbiasedness of the estimators. At $n = 50$, absolute bias does not exceed 0.012, and at $n \geq 200$ bias is effectively zero to three decimal places.

Second, RMSE decreases at the expected rate proportional to $\frac{1}{\sqrt{n}}$, confirming that the estimators are consistent. At $n = 500$, RMSE is below 0.022 in all scenarios, indicating high precision.

Third, empirical 95% confidence interval coverage is close to the nominal 95% level across all scenarios, with slight under coverage at $n = 50$ (around 93–94%). Coverage approaches the nominal level rapidly as n increases, reaching 94.7–95.2% at $n \geq 200$. The slight under coverage at small sample sizes is consistent with the use of asymptotic normal approximations, and the bootstrap confidence interval is recommended as an alternative when $n < 100$.

Fourth, estimation accuracy improves with higher prevalence levels. This reflects the fact that at very low prevalence, the number of true positives in the sample (n_{11}) is small, leading to greater variability in the conditional estimates. Researchers working with rare conditions should use larger sample sizes to ensure adequate precision.

Illustrative Example: Prostate Cancer Screening

Study Description

To illustrate the proposed method, we apply it to a real dataset from a prostate cancer screening study. A clinician collected a random sample of 135 adult male subjects from a population, of whom $n_1 = 18$ are known or believed to actually have prostate cancer and $n_2 = 117$ are known or believed not to have it, giving a total sample of n = 135. The research interest is to confirm through a diagnostic screening test whether or not each subject has prostate cancer. The reported population prevalence of prostate cancer is P(B) = 0.0005 (5 per 10,000 adult males). Consistent with the motivation for the proposed method, all estimations are first carried out assuming this prevalence is unknown, and the results are subsequently compared with those obtained using the known prevalence rate.

The results of the screening test are presented in Table II below.

Table II. Results of Prostate Cancer Screening Test of Adult Males in a Population.

Screening Test Result	**Prostate Cancer Present (B)**	**Prostate Cancer Absent ($\bar{B}$)**	**Total**
Test Positive (A)	$6(n_{11} = f^{+})$	$3(n_{12} = f^{0+} = n_{.2} - n_{22})$	$9(n_{1.})$
Test Negative ($\bar{A}$)	$12(n_{21} = f^{0-} = n_{.1} - n_{11})$	$114(n_{22} = f^{-})$	$126(n_{2.})$
$Total(n_{.j})$	$18(n_{.1})$	$117(n_{.2})$	$135(n_{..} = n)$

Computation of Frequencies and Probability Estimates

From Table II, the directly observed quantities are: $n_{11} = 6, n_{22} = 114, n_1 = 18, n_2 = 117,$

$n = 135$

The unobserved off-diagonal frequencies are estimated as:

$n_{12} =$ 117 − 114 = 3 (False Positives) n_{21} =18 – 61=2 (False Negatives)

Joint probability estimates from Equations (6a)–(6e):

$$\hat{\pi}^{+} = \frac{6}{135} = 0.0444, \qquad \hat{\pi}^{-} = \frac{114}{135} = 0.8444, \quad \hat{\pi}^{0} = \frac{15}{135} = 0.1111$$

$$\pi^{0+} = \frac{3}{135} = 0.0222(\text{joint proportion of False Positives})$$

$$\pi^{0-} = \frac{12}{135} = 0.0889(\text{joint proportion of False Negatives})$$

Concordance index and variance from Equations (7)–(8):

$$\hat{\omega} = 0.0444 - 0.8444 = -0.8000$$

$$\widehat{Var}(\hat{\omega}) = (0.0444 + 0.8444) - (-0.8000)^2 = 0.8889 - 0.6400 = 0.2489$$

$$SE(\hat{\omega}) = 0.0429$$

$$95\%\ CI\ for\ \omega: [-0.8000 - 1.96(0.0429), -0.8000 + 1.96(0.0429)] = [-0.884, -0.716]$$

Marginal proportions from Equations (9b) and (9d):

$\hat{P}(A) = (6 + 117 - 114) / 135 = 9 / 135 = 0.0667$ (6.67%) [marginal proportion]

$\hat{P}(\bar{A}) = (18 - 6 + 114) / 135 = 126 / 135 = 0.9333$ (93.33%) [marginal proportion]

Conditional Diagnostic Indices

The conditional diagnostic indices are computed from Equations (10)–(17(a)). Table V summarises all results together with 95% confidence intervals.

$$\widehat{FPR} = \hat{P}(\bar{B}|A) = \frac{3}{9} = 0.3333\ [95\%\ CI: = (0.025, 0.641)]$$

$$\widehat{TPR} = \hat{P}(B|A) = \frac{6}{9} = 0.6667\ [95\%\ CI: = (0.359, 0.975)]$$

$$\widehat{FNR} = \hat{P}(B|\bar{A}) = \frac{12}{126} = 0.0952\ [95\%\ CI: = (0.047, 0.144)]$$

$$\widehat{TNR} = \hat{P}(\bar{B}|\bar{A}) = \frac{114}{126} = 0.9048\ [95\%\ CI: = (0.856, 0.953)]$$

$$\widehat{Se} = \hat{P}(A|B) = \frac{6}{18} = 0.3333\ [95\%\ CI: = (0.116, 0.551)],$$

$$\widehat{Sp} = \hat{P}(\bar{A}|\bar{B}) = \frac{114}{117} = 0.9744\ [95\%\ CI: = (0.944, 1.000)]$$

Note on sample size: The relatively wide confidence intervals for $\widehat{FPR}$ *and* $\widehat{TPR}$ *reflect the small effective denominator* $n_{1\cdot} = \hat{P}(A) \times n = 9$ *test-positive subjects. The simulation study (Section 5) confirms that larger samples are needed for precise conditional estimation, especially in low prevalence settings.*

Interpretation of Results

The estimated $\widehat{FPR} = \hat{P}(\bar{B}|A) = 33.33\%$ indicates that, conditional on a positive test result, approximately one in three subjects is falsely positive (they do not actually have prostate cancer). Equivalently, $\widehat{TPR} = \hat{P}(B|A) = 66.67\%$ means that approximately two in three test-positive subjects are genuinely diseased. These are post-test conditional probabilities, analogous to the PPV and 1−FDR in the clinical literature.

The estimated $\widehat{FNR}$ = $\hat{P}(B|\bar{A})$ = 9.52% indicates that, conditional on a negative test result, approximately 9.52% of subjects in fact have prostate cancer. The $\widehat{TNR}$ = $\hat{P}(\bar{B}|\bar{A})$ = 90.48% confirms that approximately 90.48% of test-negative subjects are correctly identified as cancer free. These are also post-test conditional probabilities, analogous to the NPV and 1−FOR.

By contrast, Se = P(A|B) = 33.33% and Sp= P($\bar{A}$|$\bar{B}$) = 97.44% are pre-test conditional probabilities conditioned on the true disease state. They characterise the intrinsic discriminatory power of the test independently of the population prevalence. The high Sp indicates that the test is highly effective at ruling out prostate cancer in truly cancer-free subjects, while the low Se indicates that it misses a majority of truly cancerous subjects. This pattern is typical of conservative screening protocols designed to minimise false positives.

The negative index value $\hat{\omega}$= −0.80 (95% CI: −0.884 to −0.716) reflects the strong predominance of True Negatives over True Positives in this sample, as expected given the rarity of the condition. The width of the confidence interval for $\hat{\omega}$ is modest, indicating reasonable precision in the concordance estimate.

Comparison With The Bayesian Method

Re-estimation Using Known Prevalence P(B) = 0.0005

The diagnostic indices are now re-estimated using the traditional Bayesian approach, which requires the known prevalence rate P(B) = 0.0005. Using the law of total probability:

$$P(A) = Se \times P(B) + (1 - Sp) \times \big(1 - P(B)\big) \quad (18)$$

Expressed in terms of the estimated $Se\ and\ Sp$:

$$\hat{P}(A)Bayes = 1 - \widehat{Sp} + \big(\widehat{Se} + \widehat{Sp} - 1\big) \times P(B) \quad (19)$$

Substituting $\widehat{Se} = 0.3333,\ \widehat{Sp} = 0.9744\ and\ P(B) = 0.0005$:

$$\hat{P}(A)\ Bayes = 0.0256 + (0.3077)(0.0005) = 0.0258(2.58\%)$$

$$\hat{P}(\bar{A})Bayes = 0.9742(97.42\%)$$

The Bayesian TPR and FPR are:

$$\widehat{TPR}\ Bayes = \widehat{Se} \times \frac{P(B)}{\hat{P}(A)\ Bayes} = \frac{(0.3333)(0.0005)}{0.0258} = 0.0065(0.65\%)$$

$$\widehat{FPR}\ Bayes = 1 - \widehat{TPR}\ Bayes = 0.9935(99.35\%)$$

Summary Comparison

Table III. Comparison of diagnostic indices estimated by the proposed method and the traditional Bayesian method. All conditional types are specified for clarity.

Index	Proposed Method	Traditional Bayesian (P(B) = 0.0005)	Agreement
$\widehat{FPR}$ (conditional on A)	0.3333 (33.33%)	0.9920 (99.20%)	No
$\widehat{TPR}$ (conditional on A)	0.6667 (66.67%)	0.0080 (0.80%)	No
$\widehat{FNR}$ (conditional on $\bar{A}$)	0.0952 (9.52%)	0.0857 (8.57%)	Close
$\widehat{TNR}$ (conditional on $\bar{A}$)	0.9048 (90.48%)	0.9143 (91.43%)	Close

$\hat{S}e$ (conditional on B)	0.3333 (33.33%)	0.3333 (33.33%)	Yes
$\hat{S}p$ (conditional on $\bar{B}$)	0.9744 (97.44%)	0.9744 (97.44%)	Yes
$\hat{P}(A)$ (marginal)	0.0667 (6.67%)	0.0258 (2.58%)	Partial
$\hat{P}(\bar{A})$ (marginal)	0.9333 (93.33%)	0.9742 (97.42%)	Partial

When the Bayesian Method Is More Appropriate

The dramatic divergence in FPR and TPR between the two methods raises an important conceptual question: which method is more appropriate, and when?

The Bayesian result, FPR ≈ 99.35% and TPR ≈ 0.65%, is not a failure of the method but rather the mathematically correct answer to a specific question: given that a randomly selected individual from the general adult male population (with prevalence P(B) = 0.0005) tests positive, what is the probability that they truly have prostate cancer? In a population where only 5 in 10,000 men have the disease, even a test with Sp = 97.44% will produce approximately 256 false positives for every 10,000 men screened, compared with only about 1.7 true positives. The positive test result is therefore overwhelmingly likely to be a false positive.

The Bayesian approach is the appropriate and indispensable method in the following scenarios:

(i) When the clinical question is about the post-test probability for an individual patient drawn from a known population, and the clinician wishes to update a prior probability based on the test result, as in the framework of Fagan (1975).

(ii) When the goal is to compare a test's predictive values across different populations with different prevalence levels, in order to understand how the test's clinical utility varies with the disease burden of the setting.

(iii) When a reliable prevalence estimate is available and the research question is explicitly about population-level screening yield, such as: of all men who will test positive in a national screening programme, what proportion will truly have prostate cancer?

When the Proposed Method Is More Appropriate

The proposed method is most appropriate in the following scenarios:

(i) When the primary goal is to characterise the intrinsic quality of the test itself, independent of the prevalence of the disease in any particular population. The proposed $\widehat{FPR} = \hat{P}(\bar{B}|A) = 33.33\%$ and $\widehat{TPR} = \hat{P}(B|A) = 66.67\%$ reflect the actual classification performance of the test in the study sample, where the disease status of each subject is known by design.

(ii) When the population prevalence P(B) is unknown, unreliable, or not applicable (for example, in a case-control study design where the proportion of diseased subjects is fixed by design and does not reflect the population prevalence).

(iii) When the research question concerns how well the test discriminates between diseased and disease-free subjects within the studied sample, without reference to external prevalence information.

It should also be noted that the divergence between the two methods is most extreme at very low prevalence, as in the prostate cancer example. At moderate prevalence levels (P(B) ≥ 0.10), the two methods converge substantially, as shown in the simulation study results (Table IV). The choice of method should therefore be guided primarily by the nature of the research question rather than by computational convenience.

Discussion

This paper has developed, formalised, and illustrated an alternative statistical method for estimating the full set of diagnostic screening test quality indices without requiring prior knowledge of the population prevalence rate P(B). A central contribution is the precise distinction between two families of diagnostic indices: (i) disease-state-conditional indices ($Se\ and\ Sp$), which are properties of the test itself and do not depend on prevalence; and (ii) test-result-conditional indices (FPR, FNR, TPR, TNR), which are properties of the test result in a specific population and do depend on prevalence when estimated via the Bayesian route. The proposed method estimates the latter family directly from observed concordance frequencies, bypassing the Bayesian dependence on prevalence.

The principal advantage of the proposed method over the traditional Bayesian approach is its complete independence from P(B). In many real-world settings, particularly in developing countries and for rare or emerging conditions, reliable prevalence data are simply unavailable (Murtagh et al., 2007). The traditional Bayesian approach cannot be applied in these settings without making strong and often unreliable prevalence assumptions.

A secondary advantage is the availability of closed-form estimators and simple delta method standard errors (Equations 12, 16, 17(b)), which make the method immediately applicable with basic statistical software. The simulation study confirms that the estimators are nearly unbiased and that the asymptotic 95% confidence intervals achieve close to nominal coverage for $n \geq 100$, with slight undercoverage at $n = 50$ for which bootstrap intervals are recommended.

The notation throughout this paper is designed to make the conditional structure of each diagnostic index explicit. $\widehat{FPR}$ is written as $\hat{P}(\bar{B}|A)$ rather than simply $1 - Se$, and $\widehat{TNR}$ is written as $\hat{P}(\bar{B}|\bar{A})$ rather than simply Sp, to maintain the distinction between conditioning on the test result and conditioning on the true disease state. This notational discipline is important for preventing the common confusion between these two families of indices.

Several limitations should be acknowledged. The method assumes perfect reference standard classification, that is, subjects are correctly labelled as disease-positive (B) or disease-negative ($\bar{B}$). When the reference standard itself has error, the estimators will be biased, and the corrections proposed by Brenner and Gefeller (1997) should be applied. The asymptotic confidence intervals undercover at small sample sizes ($n < 100$), and the bootstrap is recommended in such cases. The proposed estimators for FPR and TPR are sensitive to the effective number of test-positive subjects, $n_{1.} = \hat{P}(A) \times n$, which can be small in rare disease settings, leading to wide confidence intervals. Finally, the framework is presented for binary test outcomes; extension to ordinal or continuous test scores through appropriate thresholding or ROC-based methods is a direction for future work.

CONCLUSION

This paper has presented a rigorous and comprehensive treatment of an alternative statistical method for estimating diagnostic screening test quality indices without requiring population prevalence data. The principal contributions are: (i) a clear and consistent probabilistic framework distinguishing disease-state-conditional indices (Se, Sp) from test-result-conditional indices (FPR, FNR, TPR, TNR); (ii) a concordance index ω with established theoretical properties; (iii) closed-form sample estimators expressed as cell frequency ratios from a 2×2 contingency table; (iv) delta method standard errors and 95% confidence intervals for all estimators; (v) a simulation study confirming near-unbiasedness, consistency, and close-to-nominal confidence interval coverage for $n \geq 100$; and (vi) a nuanced comparative analysis with the Bayesian approach, clarifying when each method is appropriate.

Applied to a real prostate cancer screening dataset ($n = 135$), the method yields $Se = 33.33\%$ (95% CI: 11.6 to 55.1%), Sp= 97.44% (95% CI: 94.4 to 100%), FPR = 33.33% (95% CI: 2.5 to 64.1%), TPR = 66.67% (95% CI: 35.9 to 97.5%), FNR = 9.52% (95% CI: 4.7 to 14.4%), and TNR = 90.48% (95% CI: 85.6 to 95.3%). The wide confidence intervals for FPR and TPR reflect the small number of test-positive subjects ($n = 9$) in this dataset, highlighting the need for larger samples when precise conditional estimation is required in low prevalence settings.

Future research should focus on: (i) bootstrap confidence intervals as alternatives to the asymptotic intervals at small sample sizes; (ii) extension to imperfect reference standards following Brenner and Gefeller (1997); (iii) generalisation to multiple diagnostic tests and multi-class outcomes; (iv) incorporation of the ω index into meta-analytic frameworks (Reitsma et al., 2005) for pooled diagnostic accuracy estimation; and (v) a full theoretical comparison with the latent class methods of Dendukuri and Joseph (2001) to establish the precise conditions under which the proposed method and the Bayesian approach yield equivalent inferences.

REFERENCES

1. Altman, D. G., & Bland, J. M. (1994a). Diagnostic tests 1: Sensitivity and specificity. British Medical Journal, 308(6943), 1552. https://doi.org/10.1136/bmj.308.6943.1552
2. Altman, D. G., & Bland, J. M. (1994b). Diagnostic tests 2: Predictive values. British Medical Journal, 309(6947), 102. https://doi.org/10.1136/bmj.309.6947.102
3. Black, W. C., & Craig, H. A. (2002). Prostate-specific antigen testing for early prostate cancer detection: Problems with the current evidence. Journal of the National Cancer Institute, 94(24), 1851–1859. https://doi.org/10.1093/jnci/94.24.1851
4. Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Statistics in Medicine, 16(9), 981–991. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N
5. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595
6. Dendukuri, N., & Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics, 57(1), 158–167. https://doi.org/10.1111/j.0006-341X.2001.00158.x
7. Fagan, T. J. (1975). Nomogram for Bayes theorem. New England Journal of Medicine, 293(5), 257. https://doi.org/10.1056/NEJM197507312930513
8. Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. Journal of Clinical Epidemiology, 56(11), 1129–1135. https://doi.org/10.1016/S0895-4356(03)00177-X
9. Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian Journal of Internal Medicine, 4(2), 627–635.
10. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747
11. Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148(3), 839–843. https://doi.org/10.1148/radiology.148.3.6867347
12. Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P., & Sterne, J. A. C. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics, 8(2), 239–251. https://doi.org/10.1093/biostatistics/kxl004
13. Hui, S. L., & Walter, S. D. (1980). Estimating the error rates of diagnostic tests. Biometrics, 36(1), 167–171. https://doi.org/10.2307/2530508
14. Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology, 141(3), 263–272. https://doi.org/10.1093/oxfordjournals.aje.a117428
15. Knottnerus, J. A., & Muris, J. W. M. (2003). Assessment of the accuracy of diagnostic tests: The cross-sectional study. Journal of Clinical Epidemiology, 56(11), 1118–1128. https://doi.org/10.1016/S0895-4356(03)00206-3
16. Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain, 8(6), 221–223. https://doi.org/10.1093/bjaceaccp/mkn041
17. Lusted, L. B. (1971). Signal detectability and medical decision-making. Science, 171(3977), 1217–1219. https://doi.org/10.1126/science.171.3977.1217

18. Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in Nuclear Medicine, 8(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2
19. Murtagh, F. E. M., Zaman, M., & Marshall, D. C. (2007). Diagnostic tests and decision-making in resource-limited settings. Bulletin of the World Health Organization, 85(4), 321–328. https://doi.org/10.2471/BLT.06.034512
20. Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. Radiology, 229(1), 3–8. https://doi.org/10.1148/radiol.2291010898
21. Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford University Press.
22. Pepe, M. S., & Janes, H. (2007). Insights into latent class analysis of diagnostic test performance. Biostatistics, 8(2), 474–484. https://doi.org/10.1093/biostatistics/kxl038
23. Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New England Journal of Medicine, 299(17), 926–930. https://doi.org/10.1056/NEJM197810262991705
24. Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology, 58(10), 982–990. https://doi.org/10.1016/j.jclinepi.2005.02.022
25. Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). Evidence-based medicine: How to practice and teach EBM (2nd ed.). Churchill Livingstone.
26. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science, 240(4857), 1285–1293. https://doi.org/10.1126/science.3287615
27. Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics, 41(4), 959–968. https://doi.org/10.2307/2530967
28. Walter, S. D., & Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. Journal of Clinical Epidemiology, 41(9), 923–937. https://doi.org/10.1016/0895-4356(88)90110-2
29. Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Reports, 62(40), 1432–1449. https://doi.org/10.2307/4586294
30. Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2002). Statistical methods in diagnostic medicine. Wiley-Interscience. https://doi.org/10.1002/0471462195
31. Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation, 115(5), 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929